Region-Based Image Classification with a Latent SVM Model

Oksana Yakhnenko, Jakob Verbeek, Cordelia Schmid

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Region-Based Image Classification with a Latent SVM Model

Oksana Yakhnenko — Jakob Verbeek — Cordelia Schmid

N° 7665

July 2011

Vision, Perception and Multimedia Understanding

ISSN 0249-6399 ISRN INRIA/RR--7665--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
GRENOBLE - RHÔNE-ALPES

Region-Based Image Classification with a Latent SVM Model

Oksana Yakhnenko , Jakob Verbeek , Cordelia Schmid

Theme : Vision, Perception and Multimedia Understanding
Perception, Cognition, Interaction
Équipes-Projets LEAR

Rapport de recherche n° 7665 — July 2011 — 13 pages

Abstract: Image classification is a challenging problem due to intra-class appearance variation, background clutter, occlusion, and photometric variability. Current state-of-the-art methods do not explicitly handle background clutter, but rely on global image representations, such as bag-of-word (BoW) models. Multiple-instance learning has been used to explicitly deal with clutter, classifying an image positively as soon as at least one image region is classified positively. In this paper, we propose a more robust latent-SVM model that, unlike multiple-instance learning, does not rely on a single image region to trigger a positive image classification. Rather, our model scores an images using all regions, and associates with each region a latent variable that indicates whether the region represents the object of interest or its background. Background and foreground regions are each scored by a different appearance model, and an additional term in the score function ensures that neighboring regions tend to take the same background/foreground label. We learn the parameters of our latent SVM model using an iterative procedure that alternates between inferring the latent variables, and updating the parameters. We compare the performance of our approach on the PASCAL VOC'07 dataset to that of SVMs trained on global BoW representations, and to a multiple-instance SVM trained on BoW representations of image regions. We show that our approach outperforms multiple-instance learning by a large margin on all classes, and outperforms global BoW models in 17 out of the 20 classes.

Key-words: image categorization, segmentation, structured support vector machines, inference

Centre de recherche INRIA Grenoble – Rhône-Alpes
655, avenue de l'Europe, 38334 Montbonnot Saint Ismier
Téléphone : +33 4 76 61 52 00 — Télécopie +33 4 76 61 52 52

Résumé :

Mots-clés :

1 Introduction and motivation

The goal of image classification is to assign a label to an image based on the presence of certain objects in the image. This is a challenging task due to intra-class variations in category appearance. For example, the class *car* contains many different types of cars, or in the case of deformable objects such as animals there is a large variation in poses. Additional appearance variations are caused by occlusions, and photometric variability due to changes in viewpoint and lighting. Image representations such as bag-of-visual-words [3], or variants based on Fisher kernels [19], have been shown to yield good performance on many challenging classification tasks. These representations aggregate the quantized image patches into a vector that represents the complete image content, *e.g*. a frequency histogram of visual words. In the aggregation process the regions are treated as an orderless set, and hence all spatial information from the image is disregarded. In an attempt to overcome this limitation, several approaches that extend the orderless representations have been proposed. Examples include the use of spatial pyramids over the image [15], or the use of attention maps to suppress background clutter [12]. However, typically these techniques are used as a pre-processing step and classify images using their global representation without attempting to combine classification with background clutter suppression explicitly.

Region-based image representations have been used frequently in the context of image annotation [2, 6, 26, 27, 28]. These approaches first segment the image into a number of regions, and then predict the labels for the individual regions. The images are then labeled with the union of the labels assigned to the regions, *i.e*. the problem is essentially treated as a multiple instance learning problem [5]. Since individual regions are classified (instead of the complete image), the clutter problem is less severe, provided that the segments correspond to object and background parts in the image. For the training images the correspondence between regions and image labels is not given, and is therefore automatically inferred. Recent part-based object detectors [9, 22] also use non-global representations, associating different subwindows of a candidate object detection window to different parts of the model. Also in this case the training data does not provide the part locations of the objects. Therefore, the part locations have to be inferred by the model both at testing and training time.

In this paper we propose a new region-based image classification approach, that differs from the related work mentioned above as follows. We use a fixed set of N image segments, and each segment is assigned to either "selected" or "de-selected" state, which, therefore, results in 2^N possible image region configurations. Like in multiple instance learning, the classification score of an image is obtained by finding the highest scoring configuration over all of the 2^N configurations of the regions. Unlike the multiple instance approach, we use both, "selected" and "de-selected" segments in order to compute a score for the image. This is similar to the part-based detectors, where two parts can be thought of as "figure" and "ground". Instead of using rectangular subwindows for our "parts", each of the image segments is either assigned to "figure" or to "ground". With each region we associate a binary latent variable indicating whether we include it for "figure" or "ground". The classification score of an image is then obtained by taking the maximum score over all possible assignment of the latent variables. For the training images we assume that only an image-level label is available. We train our models using an iterative two-stage procedure reminiscent of the Expectation Maximization algorithm. In the first step, we efficiently determine the latent variable configuration that yields the maximum score. In the second step, we update the model parameters using SVM training procedures.

With experiments on the challenging PASCAL 2007 data set we evaluate our approach, and compare it to a baseline using SVMs trained on image-level bag-of-visual-words representations as well as multiple instance SVM. Our results show that our model improves mean average precision (mAP) for all classes as compared to multiple-instance learning, and on 17 of the 20 classes as compared to the global bag-of-words baseline.

2 Related work

In this section we describe related work in image annotation, multiple-instance learning, and part-based models for object detection. We highlight the similarity between these in terms of the resulting optimization problems.

For image annotation both generative [2, 6] and discriminative [26, 27, 28] region-based models have been proposed. The idea to segment the image into a number of regions, and to associate the image labels with these regions is particularly appealing in the case of multi-label problems where each image can carry a number of labels, *e.g*. indicating the presence of several object categories. Typically, the training images are only labeled at an image level, and the correspondence between labels and regions therefore has to be established automatically by the training procedure. Generative models define a joint distribution over visual features extracted from the regions and the image labels. For example, [2] uses topic models, where each topic represents a Gaussian distribution over visual features and a multinomial distribution over image labels.

In discriminative models, *e.g*. [26, 27, 28], multiple instance learning has been used. Multiple instance learning [5] is a learning paradigm where the training data examples come in "bags" of instances, and labels are provided at the bag-level with no knowledge about instance-level labels. Negative bags contain only negative instances, while positive bags contain at least one positive instance (however it is not known which one(s)). This framework applies very naturally to image classification, where various regions of an image segmentation are modeled as the instances in the bag. An image is then labeled positively if at least one of the image regions contain the object of interest.

Both generative and discriminative region-based methods have been trained using iterative procedures that alternate between updating the model parameters, and inferring the associations between segments and labels. Performance of such methods clearly depends on the quality of the image segmentation, *i.e*. to which extent the image segments corresponds to objects in the scene. Increasing the number of segments by using multiple image segmentations has been used to increase the chances of having good segments, *e.g*. by varying the parameters of the segmentation algorithm [21]. In [17] the set regions was defined as the set of all rectangular sub-windows of an image, in combination with linear classifiers over bag-of-words appearance representations. In order to efficiently infer the region which is most likely to be positively classified they used a branch-and-bound technique [14].

Recent state-of-the-art part-based object detectors also have relied on region-based representations [9, 22]. In this case the objects have been modeled as a constellation of parts. In addition to scoring the appearance of the parts, their spatial constellation has been also taken into account by the scoring function. The training data is weakly-labeled in that no parts information (such as their location) is known, and only bounding boxes for the object instances are given. These models have been learned

by iterative techniques, that alternate an inference step in which regions of the object window are associated with the parts, and updating the models given part locations.

In essence, the same optimization problem is solved in (i) the multiple instance learning technique [1], (ii) the image classification model of [17], and (iii) the part-based detector of [9]. In each case a latent variable z indexes either (i) instances in the bag, (ii) sub-windows of an image, or (iii) part locations of an object. The object x (a bag of instances, an image, or a potential object bounding-box) is then scored by the taking the maximum over the latent variable z of a score that is computed from the features that depend on both x and z. Therefore, negative examples (bags, images, or object windows) should score low for all possible z to ensure negative classification, while for positive examples it suffices that one value of z scores high enough to trigger a positive classification.

Our model also fits in this framework and is trained with similar optimization techniques. As compared to [9], we apply our model to image classification instead of object detection, and instead of localized parts we use "distributed parts" composed of all image regions that are in the same state. As compared to [17], we use all image regions to score the image, not just the part of the image in a sub-window. By separating the image into a "figure" and "ground" regions, we can score both an object and its context and use different score functions for them when appropriate.

3 Region-based latent SVM image classification

In this section we describe in detail our region-based latent SVM image classification model. We first describe the model in Section 3.1, and in Section 3.2 we discuss how to learn the model from labeled training images. Finally, in Section 3.3, we show how to efficiently infer the latent variables during learning and to classify new images.

3.1 Feature function

Our model is based on learning weights in order to assign different scores for different regions of an image, depending on whether regions are considered as "figure" or "ground" by the model. For example in object classification, we would like to use different functions to score the object and the context in which it appears. Therefore, we segment an image into a set of N regions denoted $X = \{x_1, \ldots, x_N\}$. The segmentation can be obtained by partitioning the image using a rectangular grid, or using a segmentation algorithm such as normalized cuts [23], or super-pixels [20]. We use $x_n \in R^D$ to denote a feature vector describing the appearance, and possibly shape, of the n-th region, *e.g.* such as a bag-of-word histogram over quantized SIFT descriptors [16], or color histograms.

Throughout this paper we consider a binary classification task, and our goal is to learn a score function that is predictive for the presence of an object category in an image. We associate a latent variable z_i with each segment x_i, and denote the set of latent variables as $Z = \{z_1, \ldots, z_N\}$. The value of $z_i \in \{1, \ldots, K\}$ indicates how the segment will be scored. While the model is defined for general $K \geq 2$, in our experiments we will focus on binary latent variables with $K = 2$. In this case we can interpret the state of z_i as indicating whether the segment is treated as figure or ground. The variables z_i are latent in the sense that their value is not observed, neither for training images, nor for test images that need to be classified.

We use $G = (V, E)$ to denote a graph defined over the set of regions X such that the vertices in V corresponds to the regions x_i, and each edge $(i, j) \in E \subseteq V \times V$ connects two regions x_i and x_j. In the case of segmentation by means of a regular grid we can use a 4-connected neighborhood, or when using a general segmentation algorithm we can connect all segments that share a boundary. Given a particular assignment of the latent variables we then define a scoring function of the image as a sum of unary and pairwise terms:

$$f(X, Z) = \sum_{i=1}^{N} s_u(z_i, x_i) + \sum_{(i,j)\in E} s_p(z_i, z_j, x_i, x_j). \tag{1}$$

The unary score $s_u(z_i, x_i)$ is given by a linear function. The weights for this linear function are determined by the value of latent variable z_i. Using a weight vector w_k associated with each state $k \in \{1, \ldots, K\}$ we define $s_u(z_i = k, x_i) = w_k^\top x_i$.

The pairwise terms can implement a preference for spatially contiguous assignments of the latent variables, *e.g.* such that the figure regions tend to be connected. For each combination of states the pairwise score is defined by scalar parameter λ_{kl}, and is modulated by a similarity $p(x_i, x_j)$ between the regions $s_p(z_i = k, z_j = l, x_i, x_j) = \lambda_{kl} p(x_i, x_j)$. The term $p(x_i, x_j)$ has the effect of attenuating the pairwise cost when the regions are very dissimilar, and it can for example be defined using a distance between x_i and x_j.

In order to clarify the linear dependence of the score function on the parameters, we rewrite Eq. (1) as

$$\begin{aligned} f(X, Z) &= \sum_{k=1}^{K} w_k^\top \sum_{i:z_i=k} x_i + \sum_{k,l=1}^{K} \lambda_{kl} \sum_{(i,j)\in E} p(x_i, x_j) [\![z_i = k, z_j = l]\!] \qquad (2) \\ &= \beta^\top \Phi(X, Z), \qquad (3) \end{aligned}$$

where $[\![\cdot]\!]$ denotes the indicator function of its argument. The vector β is formed by concatenating all weight vectors w_k, and the pairwise parameters λ_{kl}. The feature function $\Phi(X, Z) = [\phi(X, Z), \psi(X, Z)]$ concatenates representations for the unary and pairwise terms. The first term of $\Phi(X, Z)$ is obtained by concatenating sums over features of regions that are assigned to same latent state $k \in \{1, \ldots, K\}$, *i.e.*

$$\phi(X, Z) = \left[\sum_{i:z_i=1} x_i, \ \ldots, \ \sum_{i:z_i=K} x_i \right]. \tag{4}$$

The second term of vector $\Phi(X, Z)$ is obtained by concatenating sums of pairwise similarities over edges, *i.e.* each entry corresponds to a combination of latent states (k, l) and contains

$$\psi(X, Z) = [\ldots \sum_{i:\ z_i=k} \ \sum_{j:\ z_j=l} p(x_i, x_j) \ \ldots]. \tag{5}$$

Now that we have defined the score function for an image X and a particular assignment of the latent variables Z, we define the score function for an image as the maximum over the scores obtained over all possible values of the latent variables:

$$f(X) = \max_Z f(X, Z). \tag{6}$$

We use the sign of the score function to classify the image, *i.e.* $y = \text{sign}(f(X)) \in \{-1, +1\}$. Thus, an image will be classified positively if at least one setting of the latent variables leads to a positive score.

3.2 Learning algorithm

To learn the model parameter vector β for a given class we follow the latent SVM framework as used in [1, 9, 17]. The learning objective is the same as defined for classical SVM training [25]: we minimize the trade-off between the hinge loss on the training images, and the ℓ_2 regularization term:

$$\min_{\beta} \ \frac{1}{2}\beta^\top \beta + C \sum_{m=1}^{M} \max\left(0, 1 - y_m f_\beta(X_m)\right), \tag{7}$$

where y_m denotes the class labels of the m-th training image X_m. The optimization problem can be equivalently written as

$$\min_{\beta, \xi_m \geq 0} \quad \frac{1}{2}\beta^\top \beta + C \sum_{m=1}^{M} \xi_m \tag{8}$$

$$\text{s.t.} \quad \exists_Z \ f_\beta(X_m, Z) \geq +1 - \xi_m, \text{ if } y_m = +1 \tag{9}$$

$$\forall_Z \ f_\beta(X_m, Z) \leq -1 + \xi_m, \text{ if } y_m = -1 \tag{10}$$

For positive examples, we require that there exists at least one configuration of the latent variables which results in predicting the positive label. For the negative examples, however, no assignment of the latent variables should result in positive prediction, *i.e.* every possible assignment of latent variables should result in negative prediction. This is a semi-convex optimization problem in the sense that the loss function is convex for negative examples (corresponding to a conjunction of linear constraints), however is not convex for positive examples (corresponding to a disjunction of linear constraints).

The model parameters β are learned using an iterative algorithm similar to Expectation Maximization [4]. In the first (expectation) step, for fixed parameters β, the latent variables of the positive examples are set to $Z_m = \arg\max_Z f_\beta(X_m, Z)$. In the second (maximization) step, we solve the optimization problem to find the best model parameters β, replacing the constraint in Eq. (9) with the stronger constraint

$$f_\beta(X_m, Z_m) \geq +1 - \xi_m, \ \text{ if } y_m = +1. \tag{11}$$

The new optimization problem, using fixed Z_m, is convex. In our implementation we use LIBLINEAR [8] to solve the convex sub-problem, but any other solver could be used, *e.g.* stochastic gradient descent as in [9].

In the convex sub-problem, there is a single constraint for each positive example. For negative examples, however, the number of constraints is exponential since $Z \in \{1, \ldots, K\}^N$. In order to handle the exponential number of constraints we use the cutting plane algorithm [11], in a similar way as it is used in structured output prediction models [24]. The cutting plane method solves the problem by maintaining a working set of active constraints. The optimization problem is solved under the constraints in the working set, and then the constraints outside of the working set are checked. If all constraints are satisfied, then the original problem has been solved. If not, the most violated constraint(s) are added to the working set and the procedure is repeated.

The final training algorithm that we use integrates the cutting plane method to handle the negative examples, and re-estimating the latent variable for the positive examples. We initialize the working set of constraints with a single setting of the latent variable for all images. We then solve for β given the constraints in the working set. For the current value of β we find for each example the highest scoring latent variable assignment $Z_m = \arg\max_Z f_\beta(X_m, Z)$. For positive examples, the corresponding constraint in the working set is replaced with Eq. (11). For negative examples, if $f_\beta(X_m, Z_m) \leq -1 + \xi_m$, is violated, then Z_m is used to add an additional constraint to the working set, otherwise nothing is done.

Note that the working set contains a single constraint for each positive image, while the number of constraints for negative images grows over the iterations. In order to limit memory requirements and to speed-up the learning procedure, it is advantageous to remove negative constraints from the working set that have been inactive for a number of iterations [9].

3.3 Inference over the latent variables

What remains to be done is to compute for each image the assignment of the latent variables Z_m that maximizes the score function: $Z_m = \arg\max_Z f_\beta(X_m, Z)$. In general, this optimization problem is an intractable NP-hard problem, and can for example be approximated by using iterative conditional modes, or loopy belief propagation. In our work, we use submodular pairwise scores, and the number states of each latent variable is $K = 2$. In this case the exact solution can be found in polynomial time using graph-cuts [13]. To ensure submodularity, we constrain $\lambda_{kk} = 0$, and $\lambda_{kl} \leq 0$ for $k \neq l$.

4 Experimental validation

We first describe our experimental setup in Section 4.1, and the experimental results in Section 4.2.

4.1 Experimental setup

Data set. The PASCAL VOC data sets have become the defacto standard to evaluate image classification experiments. The main challenge of these data sets is the wide variety of object categories, as well as the significant intra-class variation due to changes in scale, color, position of the objects, background clutter, partial visibility, and inclusion of deformable objects such as people and animals. We use the publicly available PASCAL VOC 2007 dataset [7] that consists of a total of 9,963 images in 20 categories. The data set is split into fixed training/validation (5011 images) and test sets (4952 images).

To evaluate the performance of the classifiers we use average precision (AP), which is defined as the average of precisions computed at the point of each of the relevant documents in the sequence of images ranked by decreasing classification score. The mean average precision is the mean of the AP computed for each class.

Segmentation. We experimented with several segmentation methods. The first method was a naive naive segmentation approach, cutting the image into 5×5 non-overlapping rectangles. The second approach was using normalized cuts [23] with 16 segments.

	aeroplane	bicycle	bird	boat	bottle	bus	car
BoW-SVM	**68.3**	39.1	33.2	59.0	14.2	34.2	64.0
miSVM	25.4	24.2	13.5	35.8	8.2	15.4	47.7
RBLSVM	66.9	**43.3**	32.4	**59.5**	16.0	39.2	**68.9**
RBLSVM (v)	67.7	41.4	**34.8**	59.4	**16.3**	**39.3**	68.7

	cat	chair	cow	d.table	dog	horse	motorbike
BoW-SVM	37.7	36.7	28.7	24.9	**32.4**	59.8	39.9
miSVM	26.6	21.2	12.3	10.1	20.9	42.4	18.5
RBLSVM	**38.0**	38.5	27.7	27.6	31.7	**66.7**	**45.8**
RBLSVM (v)	37.3	**39.5**	**29.5**	**28.7**	31.1	64.7	43.8

	person	p.plant	sheep	sofa	train	tvmonitor	mean AP
BoW-SVM	73.3	11.6	**30.7**	25.6	58.4	32.6	40.2
miSVM	65.9	7.6	9.3	12.8	29.2	16.5	23.2
RBLSVM	**77.0**	12.5	28.8	28.5	61.1	**35.0**	**42.3**
RBLSVM (v)	76.8	**13.3**	25.6	**29.7**	**61.8**	34.5	42.2

Table 1: Average precision (AP) for global bag-of-words SVM, region-based miSVM, and our region-based latent SVM models. Last column gives the mean AP over all classes.

Surprisingly, both segmentations produced similar performance, and thus in the remainder of the paper we use the simple grid segmentation since it is performed trivially. We use a 4-neighborhood to connect the regions in a graph.

Feature extraction. We use bag-of-visual-word representations based on quantized SIFT descriptors [16]. From each image we sampled 1000 SIFT features computed at random locations and scales, and we learned a codebook of $W = 1000$ visual words using k-means. Furthermore, we apply power-normalization to each histogram [19] by taking the square-root of the visual word histogram entries.

Pairwise potentials. Given an edge $(i, j) \in E$ in the graph we compute a pairwise similarity $p(x_i, x_j)$ as proposed in [10] given by $p(x_i, x_j) = \frac{1}{1+\|x_i - x_j\|}$ where x_i and x_j are LUV color histograms from regions i and j respectively, and $\|\cdot\|$ is the ℓ_2-norm.

4.2 Experimental results

We now proceed to compare our region-based latent SVM (RBLSVM) with two baseline approaches. The first is a (linear) SVM trained on bag-of-visual-words extracted from the entire image (BoW-SVM). The second is an SVM trained with the multiple instance learning method of [1] using each image as a bag that contains the 5×5 regions as instances (miSVM). For all models we use the validation set to set the regularization parameter C. The models are then retrained on the union of training and validation sets with the selected regularization parameter. We use the baseline BoW-SVM model trained on global image histograms to initialize the latent variables for the latent SVM model. The regions, in both positive and negative images, are then initialized by the class predicted for that region using the baseline BoW-SVM. We also include a variant of our model, RBLSVM (v), where we set the pairwise parameter using cross-validation, see below.

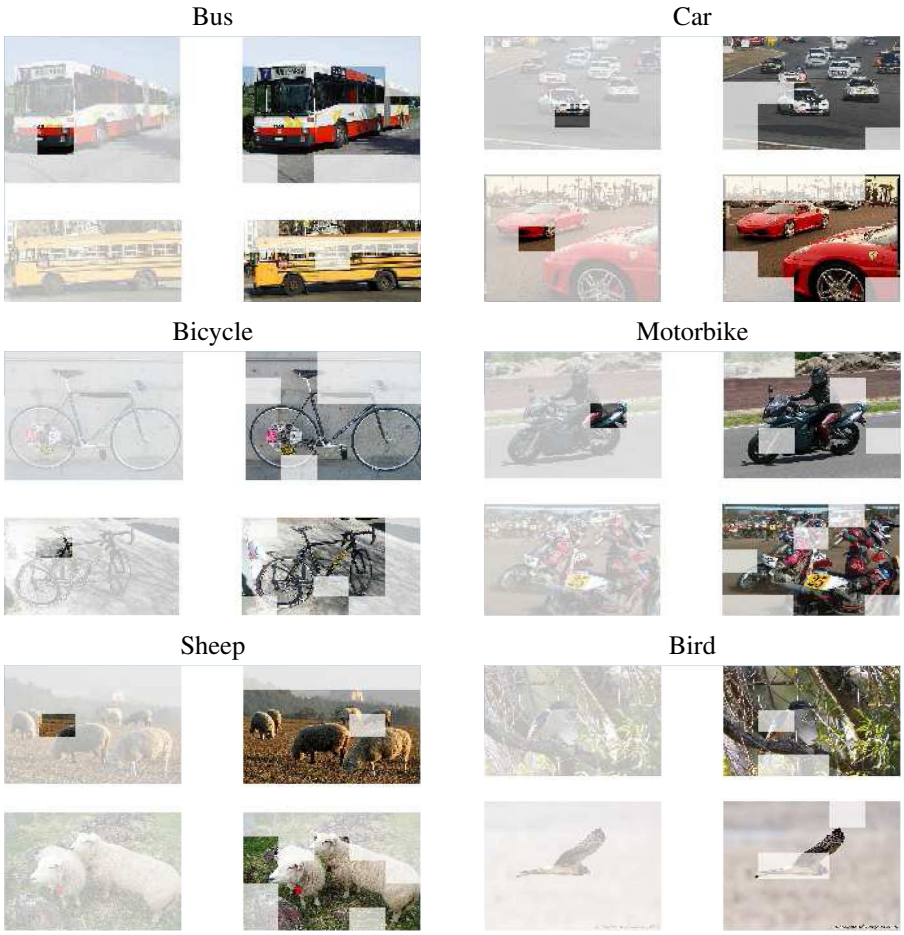

Figure 1: Examples of region classification by miSVM (left), and our RBLSVM (right). For visualization purposes, regions labeled as negative are marked by a transparent white mask.

The per-category AP and over-all performance of the four models we evaluated is reported in Table 1. First of all, it is striking that using miSVM we obtain results that are well below those obtained with the other models. This is probably due to the fact that miSVM scores the image based on the features of a single region, and therefore can not use the complete image content for classification.

Our RBLSVM improves the baseline BoW-SVM results by approximately 2% mAP, from 40.2% to 42.3%. This may appear as a small improvement, however if we take a closer look at the per-category performance we see that RBLSVM model outperforms the baseline BoW-SVM on 17 out of 20 categories. Categories of vehicles appearing in urban street scenes, *bicycle, bus, car, motorbike*, standout by particularly large improvements. This might be explained by the fact that images of these classes are easily be confused by SVMs based on image-wide bag-of-word histograms, since the global scene layout is similar for these categories. In this case the RBLSVM model can bring improvements by scoring differently the object itself and its scene context.

In Figure 1 we show the region assignments obtained using our RBLSVM model for two images of several classes. In the same figure we also show the region classification as produced by miSVM. For the latter we see that most regions are classified negatively, which could be due to the fact that miSVM learns the model under the assumption that only a single region is responsible for a positive class label. Consequently it finds solutions in which many regions score poorly and get classified negatively, and only a few regions get classified positively. Our RBLSVM model, on the other hand, generally does select the object regions in a more-or-less consistent manner, except perhaps for the *bird* category; for this class our model also performs worse than the baseline SVM.

When we examined the learned pairwise parameters λ_{kl} we found that the learned values are so small that they bring a negligible contribution to the total score function. Since the learning algorithm is not guaranteed to find the globally optimal parameters, we performed an additional experiment where we set the λ_{kl} by maximizing mAP on the validation set. The results obtained in this manner are indicated in Table 1 by RBLSVM (v). Perhaps surprisingly, setting the λ_{kl} using the validation set does not significantly change overall performance, but it does change for some of the classes.

5 Conclusion

We have presented a new region-based image classification approach that automatically decomposes images into parts that play a different role in the classification process, *e.g*. corresponding to an object of interest and its context. The model is trained in a weakly-supervised manner, *i.e*. without requiring that objects in training images are localized by bounding boxes or otherwise. The optimization problem to learn the model parameters is formulated as a latent SVM problem, which we solve using an EM-like procedure that combines an efficient solution to assign the latent variables and a cutting plane technique to efficiently handle the exponential number of constraints. We show that the proposed model outperforms a bag-of-visual-words SVM baseline on the challenging PASCAL VOC'07 data set. We also compare to a multiple instance SVM learning approach using the same image regions and features, which leads to much poorer results. This demonstrates the benefits of combining regions as done in our model, instead of selecting a single region to score the image as done in multiple instance learning, or using global bag-of-words image representations.

Several extensions to the proposed model can be envisioned. First, in our experiments we have used binary latent variables. An extension to latent variables with more states enriches the model, but also renders exact inference intractable. It is therefore unclear whether the advantage of more states outweighs the errors introduced in the inference. Second, in this work we used linear classifiers, since this requires storage requirement linear in the number of regions. This formulation allows us to use fast SVM solvers that solve the optimization problem in the primal. It is interesting to extend this approach to using non-linear kernels, which are known to yield better classification performance. Alternatively, (approximate) explicit embeddings of the induced feature spaces of non-linear kernels can be used to maintain tractable space and time requirements. Yet another option would be to use Fisher kernel image representation [18] which are also known to yield excellent performance when combined with linear classifiers.

References

[1] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems*, 2003.

[2] D. Blei and M. Jordan. Modeling annotated data. In *ACM SIGIR Conference on Research and Development in Information Retrieval*, 2003.

[3] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *ECCV Int. Workshop on Stat. Learning in Computer Vision*, 2004.

[4] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[5] T. Dietterich, R. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89:31–71, 1997.

[6] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *European Conference on Computer Vision*, 2002.

[7] M. Everingham, L. van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html, 2007.

[8] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, pages 1871–1874, 2008.

[9] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[10] B. Fulkerson, A. Vedaldi, and S. Soatto. Class segmentation and object localization with superpixel neighborhoods. In *International Conference on Computer Vision*, 2009.

[11] J. Kelley. The cutting-plane method for solving convex programs. *Journal of the Society for Industrial Applied Mathematics*, 8:703–712, 1960.

[12] F. Shahbaz Khan, J. van de Weijer, and M. Vanrell. Top-down color attention for object recognition. In *International Conference on Computer Vision*, 2009.

[13] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? *IEEE Pattern Analysis and Machine Intelligence*, 26, 2004.

[14] C. Lampert, M. Blaschko, and T. Hofmann. Efficient subwindow search: A branch and bound framework for object localization. *IEEE Pattern Analysis and Machine Intelligence*, 31(12):2129–2142, 2009.

[15] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *International Conference on Computer Vision and Pattern Recognition*, 2006.

[16] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[17] M. Nguyen, L. Torresani, F. de la Torre, and C. Rother. Weakly supervised discriminative localization and classification: a joint learning process. In *International Conference on Computer Vision*, 2009.

[18] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *International Conference on Computer Vision and Pattern Recognition*, 2007.

[19] F. Perronnin, J. Sanchez, and T. Mensink. Improving the Fisher kernel for large-scale image classification. In *European Conference on Computer Vision*, 2010.

[20] X. Ren and J. Malik. Learning a classification model for segmentation. In *International Conference on Computer Vision and Pattern Recognition*, 2003.

[21] B. Russell, W. Freeman, A. Efros, J. Sivic, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *International Conference on Computer Vision and Pattern Recognition*, 2006.

[22] P. Schnitzspan, S. Roth, and B. Schiele. Automatic discovery of meaningful object parts with latent CRFs. In *International Conference on Computer Vision and Pattern Recognition*, 2010.

[23] J. Shi and J. Malik. Normalized cuts and image segmentation. In *International Conference on Computer Vision and Pattern Recognition*, 1997.

[24] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research*, 6:1453–1484, 2005.

[25] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

[26] Z.-J. Zha, X.-S. Hua, T. Mei, J. Wang, G.-J. Qi, and Z. Wang. Joint multi-label multi-instance learning for image classification. In *International Conference on Computer Vision and Pattern Recognition*, 2008.

[27] M.-L. Zhang and Z.-H. Zhou. Multi-instance multi-label learning with application to scene classification. In *Advances in Neural Information Processing Systems*, 2006.

[28] M.-L. Zhang and Z.-H. Zhou. M3MIML: A maximum margin method for multi-instance multi-label learning. In *International Conference on Data Mining*, 2008.

Centre de recherche INRIA Grenoble – Rhône-Alpes
655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Saclay – Île-de-France : Parc Orsay Université - ZAC des Vignes : 4, rue Jacques Monod - 91893 Orsay Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399